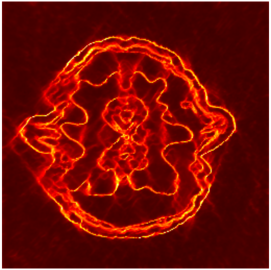
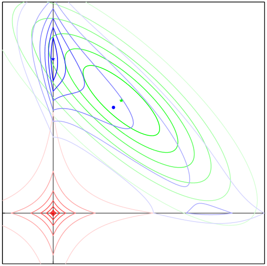
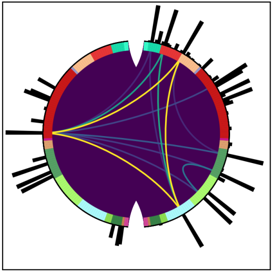
Sparse Bayesian Inference & Uncertainty Quantification for Inverse Imaging Problems







Felix Lucka
Centrum Wiskunde & Informatica
University College London
Felix.Lucka@cwi.nl

Statistics for Structures Seminar
Leiden

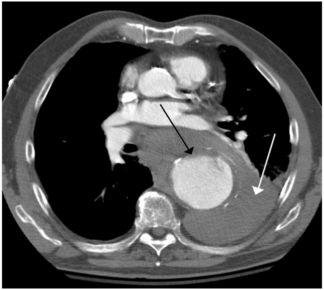
October 20, 2017

Centre for Medical Image Computing

Big Picture: From Qualitative to Quantitative Imaging

Traditional task: Produce results to be interpreted by trained experts
$\Longrightarrow$ *Qualitative* usage of the reconstructed information.

Example: Conventional *computer tomography* (*CT*).



Source: Wikimedia Commons

Big Picture: From Qualitative to Quantitative Imaging

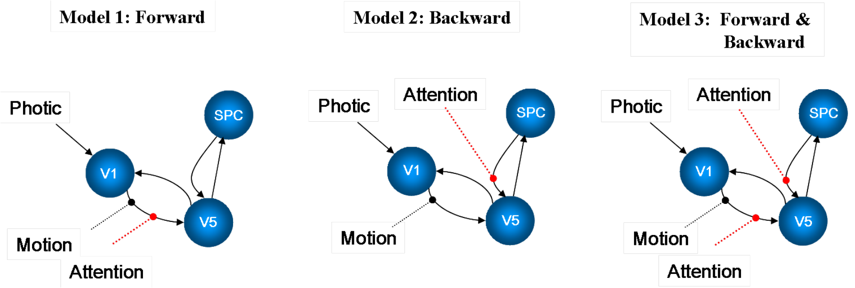
Traditional task: Produce results to be interpreted by trained experts

$\Longrightarrow$ *Qualitative* usage of the reconstructed information.

New demand: Produce results for automatized analysis procedures / hypothesis testing; multimodal imaging.

$\Longrightarrow$ *Quantitative* usage of the reconstructed information.

Example: *Dynamical causal modeling* (*DCM*).



Source: Andre C. Marreiros et al. (2010), Scholarpedia, 5(7):9568.

Bayesian Inversion and Uncertainty Quantification

Noisy, ill-posed inverse problems:

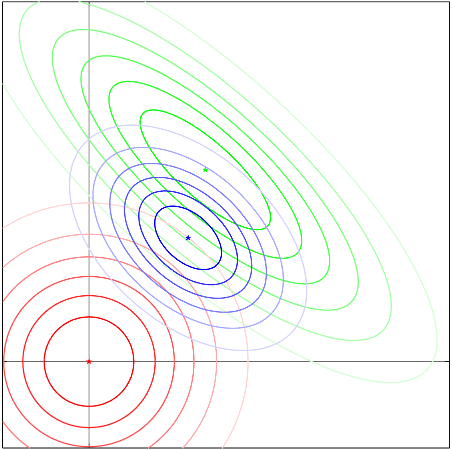
$$f = N\left(\mathcal{A}(u), \varepsilon\right)$$

Example: $f = Au + \varepsilon$

$p_{like}(f|u) \propto$
$\exp\left(-\frac{1}{2}\|f - A\,u\|_2^2\right)$

$p_{prior}(u) \propto$
$\exp\left(-\lambda\,\|D^T u\|_2^2\right)$

$p_{post}(u|f) \propto$
$\exp\left(-\frac{1}{2}\|f - A\,u\|_2^2 - \lambda\,\|D^T u\|_2^2\right)$



Probabilistic representation allows for rigorous **quantification of solution's uncertainties**.

Bayesian Inversion and Uncertainty Quantification

Noisy, ill-posed inverse problems:

$$f = N\left(\mathcal{A}(u), \varepsilon\right)$$

Example: $f = Au + \varepsilon$

$p_{like}(f|u) \propto$

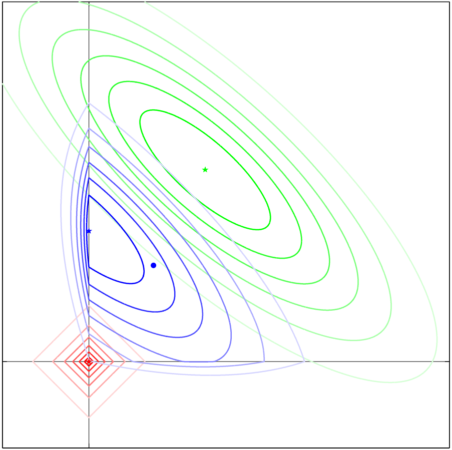
$\exp\left(-\frac{1}{2}\|f - A\,u\|_2^2\right)$

$p_{prior}(u) \propto$

$\exp\left(-\lambda\,\|D^T u\|_1\right)$

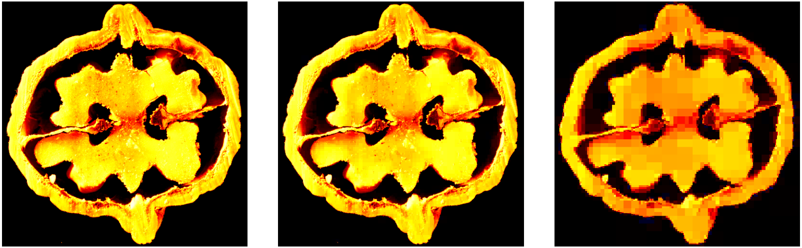
$p_{post}(u|f) \propto$

$\exp\left(-\frac{1}{2}\|f - A\,u\|_2^2 - \lambda\,\|D^T u\|_1\right)$



Probabilistic representation allows for rigorous **quantification of solution's uncertainties**.

Sparsity / Compressible Representation



(a) 100% (b) 10% (c) 1%

Sparsity as a-priori constraints are used in variational regularization, compressed sensing and variable selection:

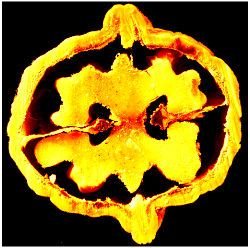
$$\hat{u}_\lambda \quad = \underset{u}{\operatorname{argmin}} \left\{ \tfrac{1}{2} \|f - A\,u\|_2^2 + \lambda \|D^T u\|_1 \right\}$$

(e.g. *total variation*, *wavelet shrinkage*, *LASSO*,...)

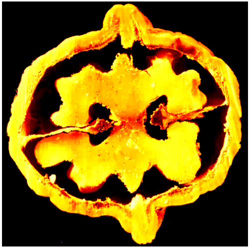
CWI

UCL

Sparsity / Compressible Representation



(a) 100%



(b) 10%

(c) 1%

Sparsity as a-priori constraints are used in variational regularization, compressed sensing and variable selection:

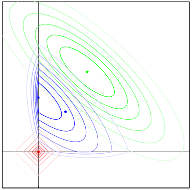
$$\hat{u}_\lambda = \operatorname*{argmin}_u \left\{ \tfrac{1}{2} \| f - A\, u \|_2^2 + \lambda \| D^T u \|_1 \right\}$$

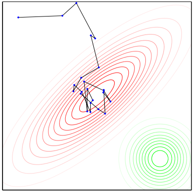
(e.g. *total variation*, *wavelet shrinkage*, *LASSO*,...)

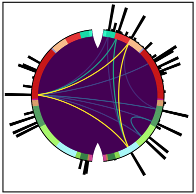
Sparse Bayesian inversion?

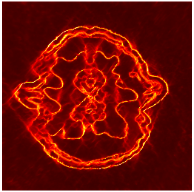
Uncertainty Quantification for Sparse Bayesian Inversion

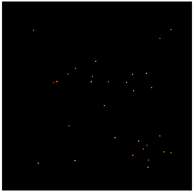
- How to model sparsity?
  - ℓ_1-norm priors.
  - Gaussian scale mixture (hierarchical Bayesian)
  - ℓ_p-norm scale mixture (hierarchical Bayesian)
- How to we compute estimators / UQ measures?
- What can we say about estimators?
- Meaningful UQ measures for sparse inversion/imaging?

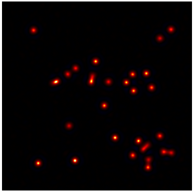








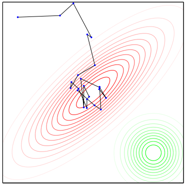




Efficient MCMC for Sparse Image Reconstruction

Task: Monte Carlo integration by samples from

$$p_{post}(u|f) \propto \exp\left(-\frac{1}{2}\|f - A\,u\|^2_{\Sigma_\varepsilon^{-1}} - \lambda\,\|D(u)\|_1\right)$$

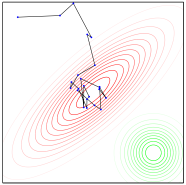


Problem: Standard Markov chain Monte Carlo (MCMC) sampler (Metropolis-Hastings) inefficient for large n or λ.

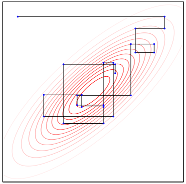
Efficient MCMC for Sparse Image Reconstruction

Task: Monte Carlo integration by samples from

$$p_{post}(u|f) \propto \exp\left(-\frac{1}{2}\|f - A\,u\|^2_{\Sigma_\varepsilon^{-1}} - \lambda\,\|D(u)\|_1\right)$$



Problem: Standard Markov chain Monte Carlo (MCMC) sampler (Metropolis-Hastings) inefficient for large n or λ.

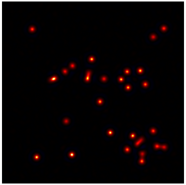


Contributions:

- Development of different Gibbs samplers.
- Efficient for high-dim. imaging ($n > 10^6$).

F.L, 2016. Fast Gibbs sampling for high-dimensional Bayesian inversion, *Inverse Problems.*

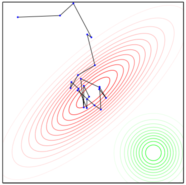
F.L, 2012. Fast Markov chain Monte Carlo sampling for sparse Bayesian inference in high-dimensional inverse problems using L1-type priors, *Inverse Problems.*



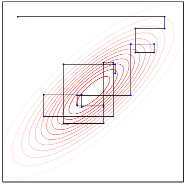
Efficient MCMC for Sparse Image Reconstruction

Task: Monte Carlo integration by samples from

$$p_{post}(u|f) \propto \exp\left(-\frac{1}{2}\|f - A\,u\|^2_{\Sigma_\varepsilon^{-1}} - \lambda\,\|D(u)\|_1\right)$$

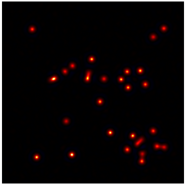


Problem: Standard Markov chain Monte Carlo (MCMC) sampler (Metropolis-Hastings) inefficient for large n or λ.



Work by Marcelo Pereyra et al.:

- Unadjusted Langevin algorithm applied to Moreau-Yoshida envelopes of posterior energy.
- As easy to implement as proximal gradient descent.



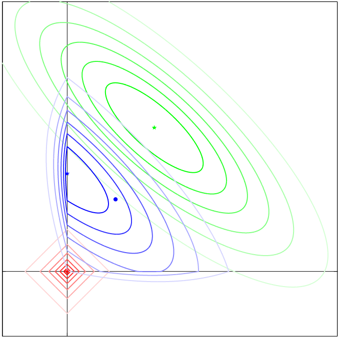
Durmus, Moulines, Pereyra, 2016. Efficient Bayesian computation by proximal Markov chain Monte Carlo: when Langevin meets Moreau, *arXiv:1612.07471.*

Point Estimators in Bayesian Inference for Imaging

$$\hat{u}_{\mathrm{MAP}} := \underset{u \in \mathbb{R}^n}{\operatorname{argmax}} \{ p_{post}(u|f) \} \quad \text{vs.} \quad \hat{u}_{\mathrm{CM}} := \int u \, p_{post}(u|f) \, \mathrm{d}u$$

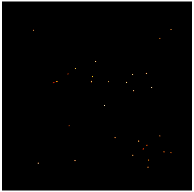
State in imaging ∼5 years ago:

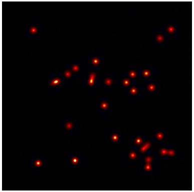
- CM preferred in theory, inaccessible in practice.
- MAP discredited by theory, accessible in practice.



Point Estimators in Bayesian Inference for Imaging

$$\hat{u}_{\text{MAP}} := \underset{u \in \mathbb{R}^n}{\operatorname{argmax}} \{\, p_{post}(u|f) \} \quad \text{vs.} \quad \hat{u}_{\text{CM}} := \int u \; p_{post}(u|f)\, \mathrm{d}u$$



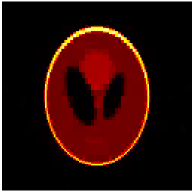


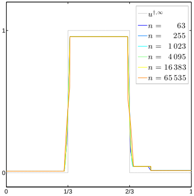
State in imaging ∼5 years ago:

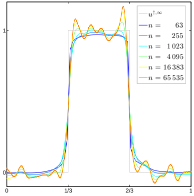
- CM preferred in theory, inaccessible in practice.
- MAP discredited by theory, accessible in practice.

However:

- MAP results looks/performs better or similar to CM.
- Gaussian priors: MAP = CM. Funny coincidence?
- Theoretical argument has a logical flaw.

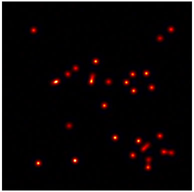






Point Estimators in Bayesian Inference for Imaging

$$\hat{u}_{\text{MAP}} := \underset{u \in \mathbb{R}^n}{\operatorname{argmax}} \left\{ p_{post}(u|f) \right\} \quad \text{vs.} \quad \hat{u}_{\text{CM}} := \int u \, p_{post}(u|f) \, \mathrm{d}u$$

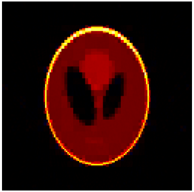


State in imaging ∼5 years ago:

- CM preferred in theory, inaccessible in practice.
- MAP discredited by theory, accessible in practice.

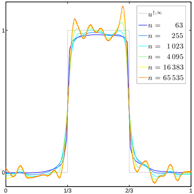
Contributions:

- Theoretical rehabilitation of MAP.
- Key: Bayes cost based on Bregman distances.
- Gaussian case consistent in this framework.

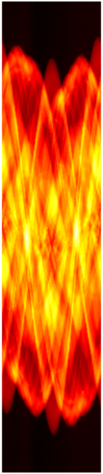


Burger & L, 2014. Maximum a posteriori estimates in linear inverse problems with log-concave priors are proper Bayes estimators, *Inverse Problems, 30(11)*.

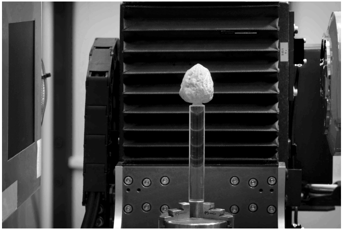
Helin & Burger, 2015. Maximum a posteriori probability estimates in infinite-dimensional Bayesian inverse problems, *Inverse Problems, 31(8)*.

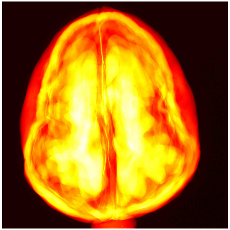


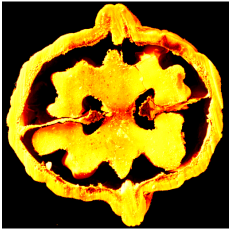
Experimental Data: Limited-Angle CT



- Cooperation with Samuli Siltanen, Esa Niemi et al.
- Besov and TV prior; non-negativity constraints.
- Stochastic noise modeling.
- Uncertainty quantification for limited angle CT.

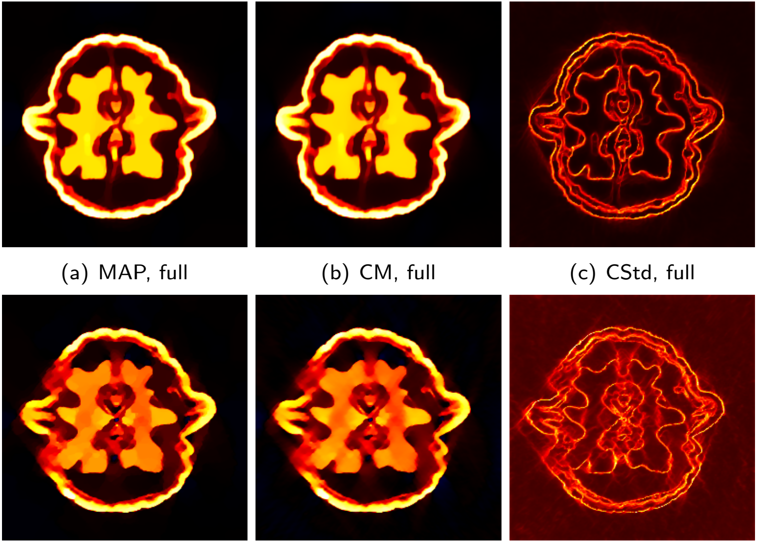






Use the data set for your own work: arXiv:1502.04064)

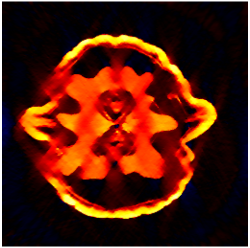
Walnut-CT with TV Prior: Full vs. Limited Angle



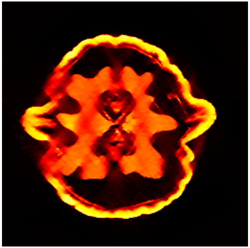
(a) MAP, full (b) CM, full (c) CStd, full

(d) MAP, limited (e) CM, limited (f) CStd, limited

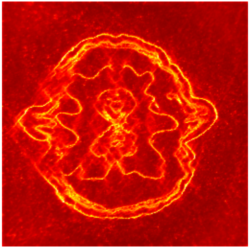
TV Prior, Non-Negativity Constraints, Limited Angle



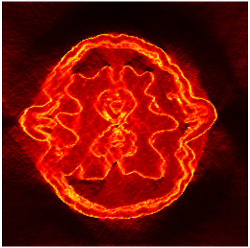
(a) CM, uncon



(b) CM, non-neg



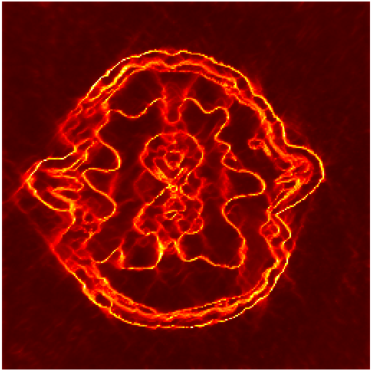
(c) CStd, uncon



(d) CStd, non-neg

However...

(a) CStd, full



(b) CStd, limited

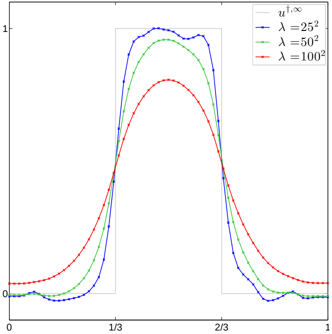
- What does it really tell me?
- Does the uncertainty decrease?!

Hierarchical Bayesian Modeling (HBM) of Sparsity

Gaussian increment prior:

$$p_{prior}(u) \propto \prod_i \exp\left(-\frac{(u_{i+1}-u_i)^2}{\gamma}\right)$$

- Gaussian variables live on characteristic scale, determined by γ.
- Similar amplitudes are likely, sparsity (= outliers) is unlikely.



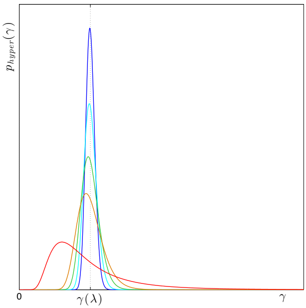
Hierarchical Bayesian Modeling (HBM) of Sparsity

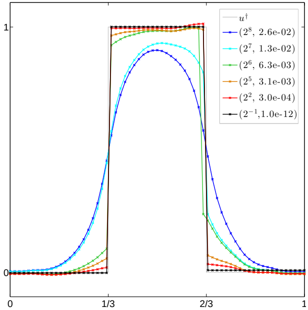
Conditionally Gaussian increment prior:

$$p_{prior}(u|\gamma) \propto \prod_i \exp\left(-\frac{(u_{i+1}-u_i)^2}{\gamma_i}\right)$$

Scale-invariant hyperprior to approximate un-informative γ_i^{-1} prior:

$$p_{hyper}(\gamma_i) \propto \gamma_i^{-(\alpha+1)} \exp\left(-\frac{\beta}{\gamma_i}\right), \quad \textit{inverse gamma distribution}$$

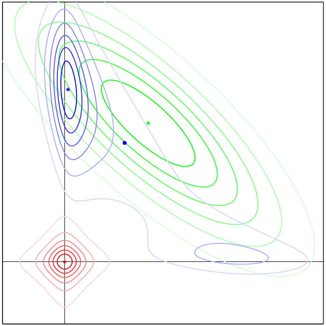


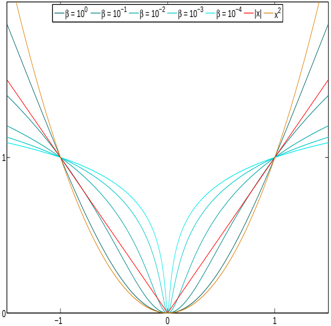


CWI

UCL

The Implicit Energy Functional behind HBM



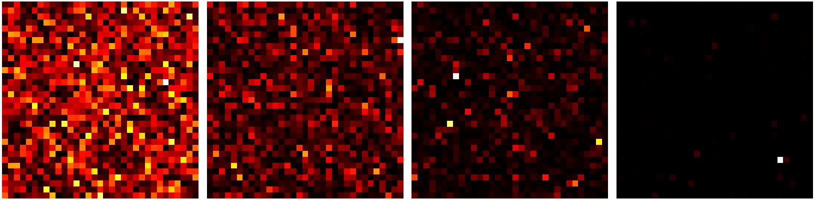


Implicit prior is a Student's t-prior with $\nu = 2\alpha, \theta = \beta/(2\alpha)$:

$$p_{prior}(u) \propto \prod_i \left(1 + \frac{u_i^2}{\nu\theta}\right)^{-\frac{\nu-1}{2}}$$

$$p_{post}(u|f) \propto \exp\left(-\tfrac{1}{2}\|f - A\,u\|^2_{\Sigma_\varepsilon^{-1}} - \tfrac{\nu-1}{2}\sum_i \log\left(1 + \frac{u_i^2}{\nu\theta}\right)\right)$$

Prior Samples



(a) ℓ_2 (b) ℓ_1 (c) $\ell_{1/2}$ (d) Cauchy

$$p_{prior}(u_i) \propto \exp(-|u_i|^p) \qquad vs. \qquad p_{prior}(u_i) \propto \frac{1}{1+u_i^2}$$

Why HBM? EEG/MEG Source Reconstruction

Aim: Reconstruction of brain activity by non-invasive measurement of induced electromagnetic fields outside of skull.

source: Wikimedia Commons

source: Wikimedia Commons

Why HBM? EEG/MEG Source Reconstruction

Aim: Reconstruction of brain activity by non-invasive measurement of induced electromagnetic fields outside of skull.

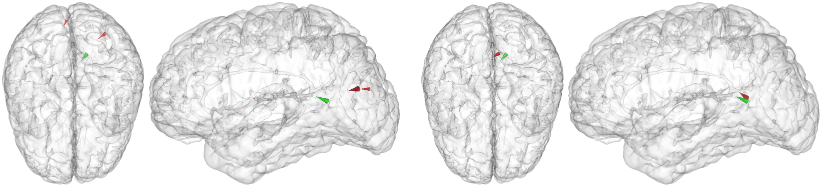
source: Wikimedia Commons

source: Wikimedia Commons

Notoriously ill-posed problem!

HBM for EEG/MEG Source Reconstruction

- Inversion with log-concave priors (e.g., ℓ_1-type) suffers from systematic depth miss-localization, HBM does not.
- HBM shows promising results for focal brain networks with simulated and real data and EEG-MEG combination.



L., Pursiainen, Burger, Wolters, 2012. Hierarchical Bayesian inference for the EEG inverse problem using realistic FE head models: Depth localization and source separation for focal primary currents, *NeuroImage, 61(4):1364–1382.*

Comparison: Two Approaches to Sparsity

feature	ℓ_p prior	HBM
$\mathcal{J}(u)$	$\|u\|_p^p$	$\frac{\nu+1}{2} \sum \log \left(1 + \frac{u^2}{\nu\theta}\right)$
sparsifying parameter	$p > 0$	$\nu > 0$
quadratic limit	$p = 2$	$\nu \rightarrow \infty$
sparse limit	$p \rightarrow 0$	$\nu \rightarrow 0$
limit functional	$\|u\|_0$	$\sum_i^n \log(\|u_i\|)$ if all $u_i \neq 0$, $-\infty$ else
solutions	sparse	compressible
differentiable	$p > 1$	always
convex	everywhere for $p \geqslant 1$	$\|u\|_\infty < \sqrt{\nu\theta}$
homogeneous	yes	no

Comparison: Two Approaches to Sparsity

feature	ℓ_p prior	HBM
$\mathcal{J}(u)$	$\|u\|_p^p$	$\frac{\nu+1}{2} \sum \log\left(1 + \frac{u^2}{\nu\theta}\right)$
sparsifying parameter	$p > 0$	$\nu > 0$
quadratic limit	$p = 2$	$\nu \to \infty$
sparse limit	$p \to 0$	$\nu \to 0$
limit functional	$\|u\|_0$	$\sum_i^n \log(\|u_i\|)$ if all $u_i \neq 0$, $-\infty$ else
solutions	sparse	compressible
differentiable	$p > 1$	always
convex	everywhere for $p \geqslant 1$	$\|u\|_\infty < \sqrt{\nu\theta}$
homogeneous	yes	no

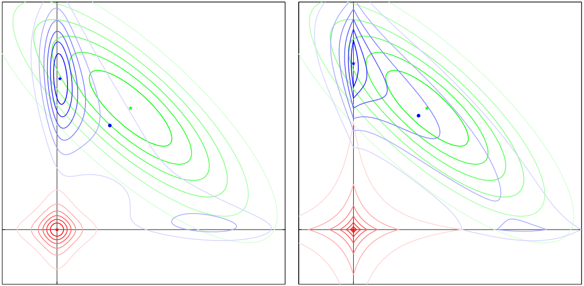
Combine them to get best (worst?!) of both worlds?

ℓ_p-hypermodels with generalized Gamma hyperpriors

$$p_{prior}(u,\gamma) \propto \exp\left(-\sum_i \left(\frac{|D_i^T u|^p}{\gamma_i} + \frac{\gamma_i^r}{\beta} - (r\alpha - 1 - 1/p)\log(\gamma_i)\right)\right)$$

Implicit prior with inverse gamma hyperprior:

$$\prod_i \left(1 + \frac{|D_i^T u|^p}{\beta}\right)^{-\alpha - 1/p}$$



(a) $p = 2$ (b) $p = 1$

ℓ_p-Hypermodels & Majorization-Minimization

Posterior with gamma hyperprior ($r = 1$), $p = 1$, and $\alpha = 2$:

$$p_{post}(u|f) \propto \exp\left(-\frac{1}{2}\|f - A\,u\|_2^2 - \sum_i \left(\frac{|D_i^T u|}{\gamma_i} + \frac{\gamma_i}{\beta}\right)\right)$$

Computational scheme for full-MAP estimation equivalent to majorization-minimization scheme for $\ell_{1/2}$ regularization (Adaptive Lasso):

$$u^{(k)} = \operatorname*{argmin}_u \left\{ \frac{1}{2}\|f - A\,u\|^2_{\Sigma_\varepsilon^{-1}} + \frac{1}{\sqrt{\beta}} \sum_i \frac{|D_i^T u|}{\sqrt{|D_i^T u|^{(k-1)}}} \right\}$$

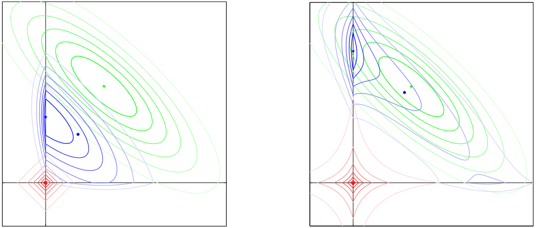
Bekhti, L, Salmon, Gramfort, 2017. A hierarchical Bayesian perspective on majorization-minimization for non-convex sparse regression: application to M/EEG source imaging, *almost submitted.*

Uncertainty Quantification for Non-Convex Sparse Recovery

Severely under-determined problems $f = Au$:

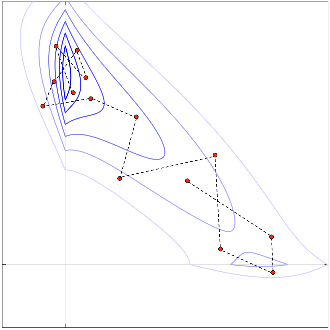
Many sparse solutions consistent with data!

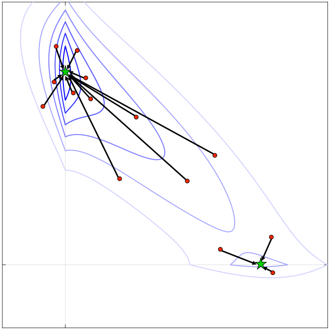
- Log-concave priors erase this ambiguity and yield single result.
- HBM posteriors get multi-modal.
- Traditional UQ measure do not capture these aspects.
- Can we preserve but quantify, structure and visualize ambiguity?



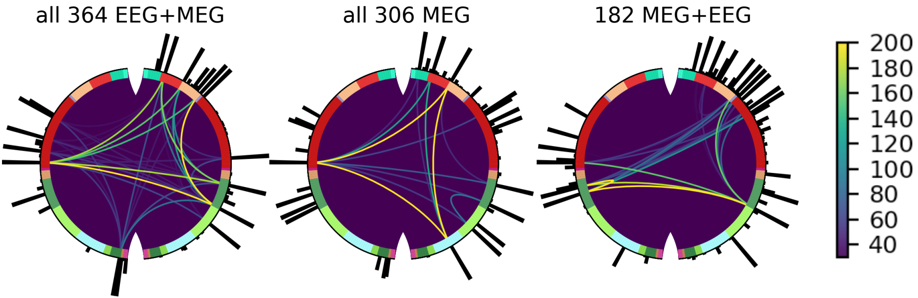
Mode Analysis with MCMC & Optimization

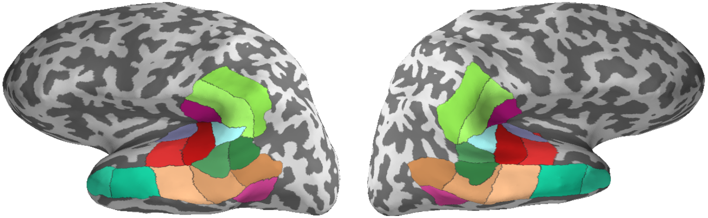
- Generate MCMC chain of posterior samples.
- Use every sample as initialization of gradient-based optimization.
- Analyse resulting chain of modes.





Sparse Source Network Analysis for EMEG Auditory Data





Summary, Outlook & Open Questions

- ℓ_p-norm and HBM road to sparsity: Neither perfect but (somewhat) computationally tractable. $\rightsquigarrow$ spike-and-slab priors?
- MAP estimates are proper Bayes estimators, modes are meaningful.
- However: Everything beyond point estimation is what's really interesting.
- Meaningful and interpretable UQ measures for sparse inversion / imaging that can complement variational approaches?
- Does it really make sense?
  (over confidence in ill-posed problems, prior domination)

- **Bekhti, L, Salmon, Gramfort, 2017.** A hierarchical Bayesian perspective on majorization-minimization for non-convex sparse regression: application to M/EEG source imaging, *almost submitted.*

- **L, 2016.** Fast Gibbs sampling for high-dimensional Bayesian inversion, *Inverse Problems.*

- **L, 2014.** Bayesian Inversion in Biomedical Imaging, *PhD Thesis, University of Münster.*

- **Burger, L, 2014.** Maximum-A-Posteriori Estimates in Linear Inverse Problems with Log-concave Priors are Proper Bayes Estimators, *Inverse Problems.*

- **L, 2012.** Fast Markov chain Monte Carlo sampling for sparse Bayesian inference in high-dimensional inverse problems using L1-type priors, *Inverse Problems.*

- **L, Pursiainen, Burger, Wolters, 2012.** Hierarchical Bayesian inference for the EEG inverse problem using realistic FE head models: Depth localization and source separation for focal primary currents, *NeuroImage.*

Thank you for your attention!

- **Bekhti, L, Salmon, Gramfort, 2017.** A hierarchical Bayesian perspective on majorization-minimization for non-convex sparse regression: application to M/EEG source imaging, *almost submitted.*
- **L, 2016.** Fast Gibbs sampling for high-dimensional Bayesian inversion, *Inverse Problems.*
- **L, 2014.** Bayesian Inversion in Biomedical Imaging, *PhD Thesis, University of Münster.*
- **Burger, L, 2014.** Maximum-A-Posteriori Estimates in Linear Inverse Problems with Log-concave Priors are Proper Bayes Estimators, *Inverse Problems.*
- **L, 2012.** Fast Markov chain Monte Carlo sampling for sparse Bayesian inference in high-dimensional inverse problems using L1-type priors, *Inverse Problems.*
- **L, Pursiainen, Burger, Wolters, 2012.** Hierarchical Bayesian inference for the EEG inverse problem using realistic FE head models: Depth localization and source separation for focal primary currents, *NeuroImage.*

MAP vs. CM Estimates: The Classical View

A theoretical argument "decides"the conflict: The Bayes cost formalism.

- An estimator is a random variable, as it relies on f and u.
- How does it perform on average? Which estimator is "best"?
- $\rightsquigarrow$ Define a cost function $\Psi(u, v)$.
- Bayes cost is the expected cost:

$$BC(\hat{u}) = \iint \Psi(u, \hat{u}(f))\, p_{like}(f|u)\, \mathrm{d}f\, p_{prior}(u)\, \mathrm{d}u$$

- Bayes estimator $\hat{u}_{BC}$ for given Ψ minimizes Bayes cost. Turns out:

$$\hat{u}_{BC}(f) = \underset{\hat{u}}{\operatorname{argmin}} \left\{ \int \Psi(u, \hat{u}(f))\, p_{post}(u|f)\, \mathrm{d}u \right\}$$

MAP vs. CM Estimates: The Classical View

Main classical arguments pro CM and contra MAP estimates:

- CM is Bayes estimator for $\Psi(u, \hat{u}) = \|u - \hat{u}\|_2^2$ (MSE).
- Also the minimum variance estimator.
- The mean value is intuitive, it is the "center of mass", the known "average".
- MAP estimate can be seen as an asymptotic Bayes estimator of

$$\Psi_\epsilon(u, \hat{u}) = \begin{cases} 0, & \text{if} \quad \|u - \hat{u}\|_\infty \leqslant \epsilon \\ 1 & \text{otherwise,} \end{cases}$$

  for $\epsilon \to 0$ (uniform cost). $\Longrightarrow$ It is not a proper Bayes estimator.

- MAP and CM seem theoretically and computationally fundamentally different $\Longrightarrow$ one should decide.
- *"A real Bayesian would not use the MAP estimate"*
- People feel "ashamed" when they have to compute MAP estimates (even when their results are good).

CWI

UCL

A False Conclusion

"A real Bayesian would not use the MAP estimate as it is not a proper Bayes estimator".

"MAP estimate can be seen as an asymptotic Bayes estimator of

$$\Psi_\epsilon(u, \hat{u}) = \begin{cases} 0, & \text{if} \quad \|u - \hat{u}\|_\infty < \epsilon \\ 1 & \text{otherwise,} \end{cases}$$

for $\epsilon \to 0$.
???$\Longrightarrow$??? It is not a proper Bayes estimator."

"MAP estimator is asymptotic Bayes estimator for some degenerate Ψ"
$\not\Rightarrow$ "MAP can't be Bayes estimator for some proper Ψ" !!!!

Two New Bayes Cost Functions

Define

(a) $\Psi_{\text{LS}}(u, \hat{u}) \; := \|A(\hat{u} - u)\|^2_{\Sigma_\varepsilon^{-1}} + \beta \|L(\hat{u} - u)\|_2^2$

(b) $\Psi_{\text{Brg}}(u, \hat{u}) := \|A(\hat{u} - u)\|^2_{\Sigma_\varepsilon^{-1}} + \lambda D_{\mathcal{J}}(\hat{u}, u)$

for a regular L and $\beta > 0$.

Properties:

- Proper, convex cost functions
- For $\mathcal{J}(u) = \beta/\lambda \|Lu\|_2^2$ (Gaussian case!) we have $\lambda D_{\mathcal{J}}(\hat{u}, u) = \beta \|L(\hat{u} - u)\|_2^2$, and $\Psi_{\text{LS}}(u, \hat{u}) = \Psi_{\text{Brg}}(u, \hat{u})$!

Theorems:

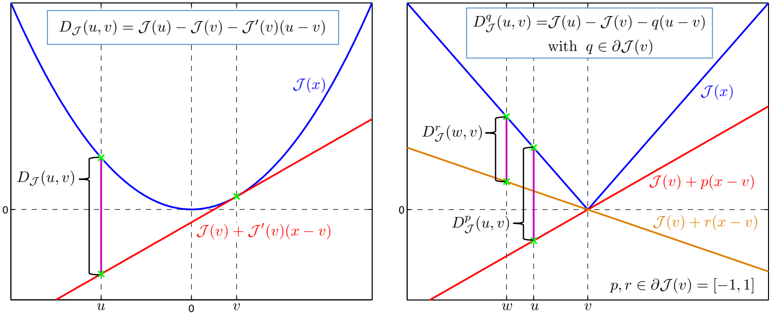
(I) The CM estimate is the Bayes estimator for $\Psi_{\text{LS}}(u, \hat{u})$

(II) The MAP estimate is the Bayes estimator for $\Psi_{\text{Brg}}(u, \hat{u})$

Bregman distances

For a proper, convex functional $\Psi : \mathbb{R}^n \longrightarrow \mathbb{R} \cup \{\infty\}$, the *Bregman distance* $D^p_\Psi(f,g)$ between $f, g \in \mathbb{R}^n$ for a subgradient $p \in \partial\Psi(g)$ is defined as

$$D^p_\Psi(f,g) = \Psi(f) - \Psi(g) - \langle p, f-g \rangle, \qquad p \in \partial\Psi(g)$$



(c) $\mathcal{J}(x) = x^2$ (d) $\mathcal{J}(x) = |x|$

Basically, $D_\Psi(f,g)$ measures the difference between Ψ and its linearization in f at another point g